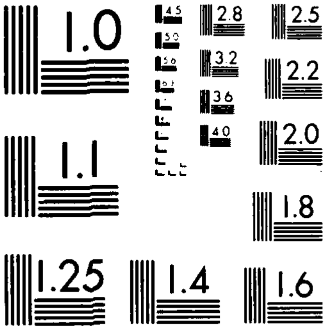
AD-A184 505   INTEGRATING SYNTAX SEMANTICS AND DISCOURSE DARPA   1/1
(DEFENSE ADVANCED RESEAR (U) UNISYS CORP PAOLI PA
CUSTOMS SYSTEMS DIV L HIRSCHMAN 12 AUG 87
UNCLASSIFIED   N00014-85-C-0012   F/G 12/5   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Custom Systems Division

UNISYS

Unisys Corporation
PO Box 517
Paoli PA 19301

Telephone
215 648 7200

HDT-87-096
12 August 1987

Office of Naval Research
Department of the Navy
800 N. Quincy Street
Arlington, VA 22217-5000

Attention: D. Allen Meyrowitz, Code 1133

Reference: DARPA Contract No. N00014-85-C-0012

Subject: Status Report of the "Darpa Natural Language Understanding Program"
Reporting Period 5/1/87 - 7/31/87

Gentlemen:

In accordance with the referenced contract requirements, we are pleased to submit an R & D Status Report for the DARPA Natural Language Understanding Program.

For any questions, please feel free to contact either Dr. Lynette Hirschman, Principal Investigator (215/648-7554) or the undersigned (215/648-2263).

Very truly yours,

Unisys, Defense Systems
(formerly System Development Corporation)

H D Tuck
H. D. Tuck
Contract Manager

HDT/d

cc: See Attached Distribution

Distribution

2 copies

LCOL Robert Simpson
IPTO
Defense Advanced Research Projects
Agency
1400 Wilson Boulevard
Arlington, VA 22209

1 copy

Defense Contract Administration Serv.
Management Area-Philadelphia
P. O. Box 7699
Philadelphia, PA 19101-7699
Attn: Mr. Al Stein, DCASR-PHI-GAAD-E1

6 copies

Director, Naval Research Laboratory
Attn: Code 2627
Washington, D.C. 20375

12 copies

Defense Technical Information Center
Bldg. 5 Cameron Station
Alexandria, Virginia 22314

INTEGRATING SYNTAX, SEMANTICS, AND DISCOURSE DARPA NATURAL LANGUAGE UNDERSTANDING PROGRAM

R&D STATUS REPORT
Unisys/Defense Systems

ARPA ORDER NUMBER: 5262
PROGRAM CODE NO. NR 049-602 dated 10 August 1984 (433)
CONTRACTOR: Unisys Defense Systems CONTRACT AMOUNT: $882,833
CONTRACT NO: N00014-85-C-0012
EFFECTIVE DATE OF CONTRACT: 4/29/85 EXPIRATION DATE OF CONTRACT: 4/28/89
PRINCIPAL INVESTIGATOR: Dr. Lynette Hirschman PHONE NO. (215) 648-7554

SHORT TITLE OF WORK: DARPA Natural Language Understanding Program

REPORTING PERIOD: - 5/1/87-7/31/87

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

Contents :

1. Description of Progress

Because the quarterly progress report for February through April was not issued separately, but was included in the Final Report for the first two years of the contract, some items from that period are repeated here for completeness.

1.1. Grammar

1.1.1. Intermediate Syntactic Representation

Rules have been added in the Intermediate Syntactic Representation (ISR) component to produce a regularized representation for verbs with multiple subcategorizations. Additional rules have been added to deal with verb-particle combinations when the verb and particle are not adjacent. A general revision of the overall ISR mechanism has been designed and is currently being tested.

1.2. Syntax/Semantics Interaction

Two changes for syntax/semantics interaction were partially implemented, but further development has been postponed until after the implementation of the newly planned Intermediate Syntactic Representation (ISR). A restriction providing a basis for the initial version of the syntax/semantics interaction mechanism was implemented which would call semantics after the subject and the main verb have both been parsed, but before the object has been analyzed. At this point, the search space for the object could be fruitfully pruned given appropriate interaction with semantics. Information returned from the semantic interpreter could be used to reorder the list of object options, or to abort the current analysis. The semantic interpreter was changed to make it possible to get access to selectional information without calling reference resolution. However, the implementation details regarding the two types of responses, i.e., to predict vs. to accept/reject, depend on the form of new ISR.

1.3. Semantics

A new meta-version of the semantics interpreter was implemented which allows it to operate in different modes, depending on the type of expression being processed. The new interpreter explicitly encodes the important similarities and differences in the processing of distinct types of predicating expressions, i.e., verbs, nominalizations, and noun predicates. A mode-setting parameter determines which optional steps in the basic algorithm need to be executed, and also selects the appropriate syntactic mapping rules.

In the past, most of the nominalizations in the corpus processed by PUNDIT have been those derived by irregular suffixes, hence, no morphological analysis was performed and the identification of the related verb was handled by special rules. Now, as a consequence of changes to the ISR and the semantics, the regular gerundive nominalizations are now processed in a fully general way which takes into account both their semantic and morphological regularity.

1.4. Work on Multiple Domains

Recent developments to PUNDIT have been funded primarily under the NSF contract, specifically for the NOSC (Naval Ocean Systems Command) Message Understanding Conference (MUCK). Similarly, changes in the environment have been funded through IR&D. Extensions and improvements that have been integrated into PUNDIT will be briefly summarized, since they pertain to future development under the DARPA contract. The most important change to the working environment consists of a new set of procedures for facilitating the concurrent development of multiple domain dependent versions of PUNDIT while insuring that the domain independent core of the system can be easily maintained and extended. These issues are especially important now that we expect to work on the RAINFORM messages from the MUCK domain concurrently with the CASREPs.

The opportunity to port the PUNDIT system to the RAINFORMs domain was a useful exercise and led to extensions in the system as a whole which will augment its coverage in any domain. Significant extensions include syntactic coverage of run-on sentences, a capacity to handle verb/particle constructions (changes to ISR and

semantic pre-processor), and semantic coverage of "transparent" verbs for which an argument of the matrix verb can be passed via special rule to fill a semantic role of the verb's nominal argument. The top level of the system was also changed to incorporate header information into the initial discourse context for pragmatic purposes.

The experience in working on the RAINFORMs concurrently with the CASREPs led to some practical refinements of the theoretical distinction between domain dependent and domain independent information. For example, the default referents used by the reference resolution component are defined in the domain model for the CASREPs domain, but come from the header for the RAINFORMs domain. A core version of the PUNDIT system has been defined and isolated from the code containing domain dependent superstructure. The tools used in creating the various executable versions of the system will be revised to reflect this distinction.

1.5. Environment

The work described here on improving our development environment has been funded under IR&D. Since it is relevant to our Darpa work, it is briefly summarized here.

1.5.1. Sun Interface

A new demo/development environment was implemented on the SUN workstations which makes use of Xwindows. This evironment provides graphical parse trees as an aid in grammar development. This required changes to PUNDIT's Prolog top level as well as new provisions for Prolog/C communication and C representation of Prolog data structures. The demo environment was used for the Pundit demo on the Sun at the ACL conference, and was judged to be very effective. Pundit ran with a new version of Quintus Prolog, Quintus Prolog 2.0, without any changes for this demo.

In order to take advantage of the new interface between PUNDIT and Xwindows, work has begun on creating a graphics display of the output of temporal analysis. A schematic layout of the graphics display has been designed, and a new Prolog data structure for representing this layout has been partially implemented.

1.5.2. Port to the Unisys Explorer

The Lisp window interface on the Explorer has now been completely implemented with full communication between the Prolog output of PUNDIT and the Lisp reader. The window interface developed for demonstrating PUNDIT on the Explorer is being rewritten to provide a transparent environment across the SUNS and the Explorer.

1.5.3. Tools

Work has progressed on two new or existing tools. A new tool has been added to remove things from the recorded database, which has a menu facility. The syntactic lexicon, the bnf definitions used by the parser, and the several categories of rules used by the semantic interpreter are all in the recorded data base. The menu makes it possible to selectively remove any one set of rules, or all of the semantics rules. Another new tool is an on-line help facility for the PUNDIT development environment.

1.6. PUNDIT Applications

Three government natural language applications have been pursued this quarter. The previous quarterly report referred to a potential internal collaboration on a contract originally held by Sperry to process messages originating on Trident submarines. The collaboration has been approved and will be funded under IR&D. A government agency RFP that describes an application for the analysis of messages about international terrorist activities into SQL database relations has has been sent out and is being evaluated. We are responding to an RFP issued by the IRS which includes a task to build a natural language interface to expert systems.

2. Change in Key Personnel

Catherine Ball, a Ph.D. candidate in linguistics at the University of Pennsylvania, and formerly of Shared Medical Systems, started working in the natural language group on April 20.

Leslie Riley, a member of the group since 1985, will be working part time while pursuing full-time studies in computer science.

Carl Weir, formerly of MCC has accepted our offer, and will be starting in September.

Shirley Steele, formerly of Bell Labs, has accepted our offer to initiate a program in speech research, and will also be starting in September.

3. Summary of Substantive Information from Meetings and Conferences

3.1. Darpa Meetings

3.1.1. Workshop on Spoken Language Systems (University of Pennsylvania)

Lynette Hirschman, Martha Palmer and Deborah Dahl attended the three-day Workshop on Spoken Language Systems, to review DARPA Strategic Computing Natural Language work and to plan for the next phase of Strategic Computing, "Spoken Language Systems". We are participating in the planning of the new program, working with Aravind Joshi at the University of Pennsylvania (see description of planning meeting held at the ACL conference, below). At the meeting, the transition of the Natural Language program from Bob Simpson to Allen Sears was announced, consolidating natural language and speech efforts.

At the Workshop, Rebecca Passonneau and Francois Lang gave demos of the PUNDIT system doing message processing and demonstrating acquisition of semantic patterns. Allen Sears attended the second demo. We also had an opportunity to review the demonstrations of other systems.

Lynette Hirschman participated in a meeting of DARPA NL contractors, held before the ACL at SRI (in Palo Alto) July 6. The goal of the meeting was to draft a white paper for Allen Sears, describing research goals for a 6-year program in Spoken Language Systems. The working group (with participants from SRI, BBN, Unisys, NYU, U. Penna and ISI) sketched out a three-year and a six-year set of objectives. A. Joshi (U. Penna) is assembling the detailed inputs and will submit the white paper to Sears at the end of July.

3.2. Professional Meetings Attended

3.2.1. MUCK: Message Understanding Conference (NOSC)

(This conference was attended under NSF funding.)

Lynette Hirschman and Rebecca Passonneau attended the Message Understanding Conference at NOSC (Naval Ocean Systems Command), San Diego, June 10-12. Two goals in participating in the conference were to find out how portable the PUNDIT system was, and to compare our system to other prototype or commercial systems. The domain of intelligence messages (RAINFORMS) chosen to demonstrate system extensibility, turned out to be quite well-suited to the capabilities of PUNDIT. We received 10 sample messages and were able to get 7 of them to run by the time of the demo. PUNDIT was the most ambitious system demonstrated, in that it is based on a principled linguistic analysis in all phases (syntax, semantics, pragmatics). Its weakness is in the limited use made of the domain model, although we plan major work on this during the second half of 1987.

The four members of the DARPA Strategic Computing text processing group (Ralph Grishman, NYU, Jerry Hobbs, SRI and the Unisys contingent) who attended the MUCK conference took the opportunity to discuss the advantages of selecting the same domain--Navy RAINFORM (intelligence) messages--for the follow-on phase of the DARPA contracts. The PUNDIT group plans to carry on development in both the CASREP and RAINFORM domains concurrently.

3.2.2. Association for Computational Linguistics:

Members of the natural language group attended the July 1987 ACL conference, where they gave demonstrations of PUNDIT (to over 50 people) and presented three technical papers. Deborah Dahl presented the paper she co-authored with Martha Palmer and Rebecca Passonneau entitled "Nominalizations in PUNDIT"; Rebecca Passonneau presented her paper describing PUNDIT's temporal analysis, "Situations and Intervals"; and Bonnie Webber presented her paper, "The Interpretation of Tense in Discourse", on research she did in collaboration with the UNISYS NL group. We demonstrated Pundit on a Sun workstation and on the Unisys Explorer to over 50 people in the course of 4 days.

3.2.3. Other Conferences and Workshops Attended

Deborah Dahl attended the In-Depth Technical Review of the MCC Natural Language groups in Austin, May 19-21.

Marcia Linebarger attended the workshop on "The Lexicon in Theoretical and Computational Perspective", Stanford, CA, July 13-24.

Catherine Ball and Francois Lang attended AAAI-87, Seattle, WA, July 12-17, where they demonstrated the PUNDIT system. The demonstration attracted many visitors and also led to some hiring contacts.

0. Problems Expected or Anticipated

None.

1. Action Required by the Government

None.

2. Fiscal Status

(1) Amount currently provided on contract:
$ 872,833 (funded) $1,704,901 (contract value)

(2) Expenditures and commitments to date:
$ 675,039

(3) Funds required to complete work:
$ 197,794

END

10-87

DTIC